

Road Detection via On-line Label Transfer

Jose M. Alvarez, Ferran Diego, Joan Serrat, Antonio M. López

Abstract—Vision-based road detection is an essential functionality for supporting advanced driver assistance systems (ADAS) such as road following and vehicle and pedestrian detection. The major challenges of road detection are dealing with shadows and lighting variations and the presence of other objects in the scene. Current road detection algorithms characterize road areas at pixel level and group pixels accordingly. However, these algorithms fail in presence of strong shadows and lighting variations. Therefore, we propose a road detection algorithm based on video alignment. The key idea of the algorithm is to exploit the similarities occurred when a vehicle follows the same trajectory more than once. In this way, road areas are learned in a first ride and then, this road knowledge is used to infer areas depicting drivable road surfaces in subsequent rides. Two different experiments are conducted to validate the proposal on different video sequences taken at different scenarios and different daytime. The former aims to perform on-line road detection. The latter aims to perform off-line road detection and is applied to automatically generate the ground-truth necessary to validate road detection algorithms. Qualitative and quantitative evaluations prove that the proposed algorithm is a valid road detection approach.

Index Terms—Road detection, image processing, video analysis, on-line video alignment.

I. INTRODUCTION

Vision-based road detection aims to detect the free road surface ahead of the ego-vehicle using an on-board camera (Fig. 1a). Road detection is a key component in autonomous driving to solve specific tasks such as road following, car collision avoidance and lane keeping [1], [2]. Moreover, it is an invaluable background segmentation stage for other functionalities such as vehicle and pedestrian detection [3]. Road detection is very challenging since the algorithm must deal with continuously changing background, the presence of different objects like vehicles and pedestrian, different road types (urban, highways, back-road) and varying ambient illumination and weather conditions (Fig. 1b).

Common vision-based road detection algorithms consider road homogeneity to group pixels according to features extracted at pixel-level such as texture [4] and color [5]. However, algorithms based on low-level features may fail for severe lighting variations (strong shadows and highlights) and may depend on structured roads. The performance of these systems is often improved by including constraints such as road shape restrictions [5] or temporal coherence [6] at the expense of limiting the applicability of the algorithm.

This work is supported by Spanish MINECO projects TRA2011-29454-C03-01, TIN2011-25606, TIN2011-29494-C03-02, and the Research Program Consolider Ingenio 2010: MIPRCV (CSD200700018) and the Catalan Generalitat project CTP-2008ITT00001..

The authors are with Computer Vision Center & Computer Science Dept., Edifici O, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Spain. jalvarez@cvc.uab.es

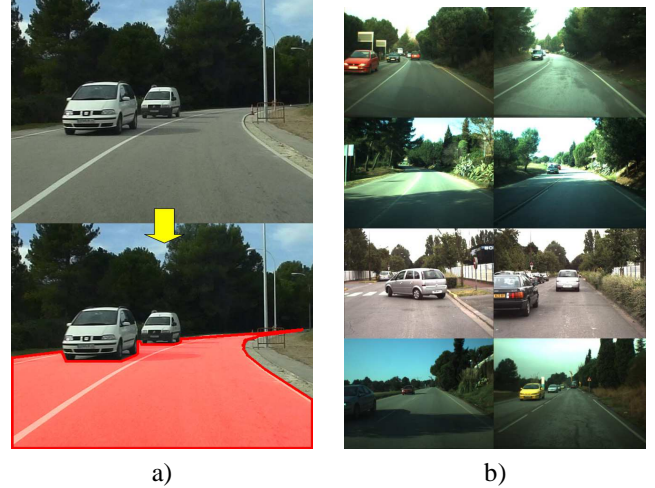


Fig. 1. (a) Vision-based road detection aims to detect the free road surface ahead a moving vehicle. (b) The main challenges of road detection are continuously changing background, the presence of different objects like vehicles and pedestrian, different road types (urban, highways, back-road) and varying ambient illumination and weather conditions.

In this paper, as a novelty, we propose a road detection approach based on video alignment. Video alignment algorithms aim to relate frames and image coordinates between two video sequences [7]. Hence, the key idea of the proposed algorithm is to exploit similarities occurred when one vehicle drives through the same route (i.e., similar trajectories) more than once (Fig. 2). In this way, road knowledge is learnt in a first ride and then, video alignment is used to detect the road in the current image by transferring this knowledge from one sequence to the current one. The result is a rough segmentation of the road that is refined to obtain the accuracy required.

The novelty of the paper is twofold: first, we propose an on-line method to perform video alignment based on image comparisons and a fixed-lag smoothing approach [8]. This method is specially designed to deal with specific road detection requirements: independent camera trajectories and independent vehicle speed variations. Second, a road detection algorithm is proposed on the basis of on-line video alignment. The algorithm improve the robustness of video-alignment to shadows by computing image comparisons in an illuminant-invariant feature space. Then, this robustness is combined with a refinement step at pixel-level to achieve the required accuracy.

The rest of this paper is organized as follows: first, in Sect. II related work is review. Then, in Sect. III, the method to perform on-line video alignment approach is outlined. The algorithm for road detection using on-line video alignment is described in Sect. IV. In Sect. V, two different experiments are presented to validate the algorithm. The former aims to

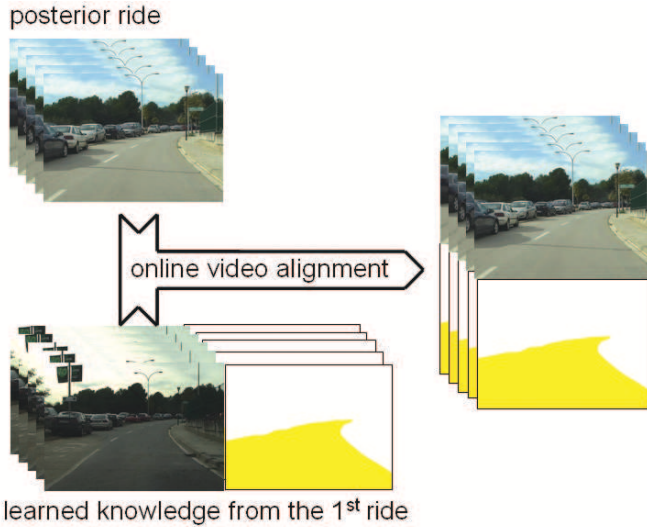


Fig. 2. Knowledge learned in a first ride is used to detect road regions in subsequent rides.

perform on-line road detection. The latter aims to perform off-line road detection and is applied to automatically generate the ground-truth necessary to validate road detection algorithms. Finally, in Sect. VI, conclusions are drawn.

II. RELATED WORK

Vision-based Road Detection. Road detection algorithms aim to detect the free road surface ahead of the ego-vehicle using an on-board camera. Common road detection algorithms consider road homogeneity to group pixels according to low-level features (pixel-level) such as texture [9], [4] and color [1], [5], [10]. For instance, in [9], Lombardi *et al.* use a textureless descriptor to characterize road areas. However, the imaged road texture varies too much with the distance to the camera due to the perspective effect. In [4], Rasmussen *et al.* use dominant orientations based on Gabor filtering to detect the vanishing point. However, this approach shows dependency on strong textures parallel to the road direction in the form of lane markings for paved roads or tracks left by other vehicles in rural (unpaved) roads. In contrast, Kong *et al.* [11] detect vanishing points using an adaptive soft-voting scheme based on confidence-weighted Gabor filters. Color appearance information has been widely accepted as the main cue for road detection since color imposes less physical restrictions (regarding the shape of the road), leading to more versatile systems. The two most popular color spaces, that have proved to be robust to minor illuminant changes, are *HSV* [5], [12] and normalized *RGB* [6]. However, algorithms based on these color spaces may fail under wide lighting variations (strong shadows and highlights among others) and these algorithms depend on highly structured roads, road homogeneity, simplified road shapes, and idealized lighting conditions. The performance of these systems is sometimes improved by including constraints such as temporal coherence [13], [6] or road shape restrictions [14].

Video Alignment. Video alignment algorithms aim to relate frames and image coordinates between two video se-

quences [15], [7], [16], [17], [18], [19]. One of these sequences is designated as observed sequence, then the other is designated as reference. The observed sequence provides the spatial and temporal reference whereas the reference one is mapped to match it. Current video alignment focus on synchronizing sequences simultaneously recorded with fixed or rigidly attached cameras. These assumptions involve a fixed spatio-temporal parametric model along the whole sequence. Hence, the video alignment is posed as a minimization problem over small amount of parameters comparing some data extracted from the images. For instance, a common approach consists of computing image similarities based solely on the gray level intensity [7], [20]. Other approaches [16], [19] exploit the benefit of temporal information and track several characteristic points along the sequences. However, all these approaches are based on rigid camera attachment and can not deal with the specific requirements of road detection: (1) independent similar trajectories and (2) independent vehicle speed variations. Another set of works address the challenge of aligning sequences recorded by independent moving cameras at different times [21], [22]. However, these algorithms require a high computational cost and they can not be applied to align sequences during acquisition. Therefore, in the next section, we propose an on-line video alignment based on a fixed-lag smoothing approach that yields an on-line video alignment estimation apart from a small fixed delay in processing the incoming data.

III. ON-LINE VIDEO ALIGNMENT

In this section, as a novelty, we propose a video-alignment approach that is able to estimate spatio-temporal relationship between two video sequences while one of them is being acquired. That is, each newly acquired in the observed sequence is mapped and pixel-wise related to one of the frames in the reference sequence. The proposed algorithm is based on [22] to deal with road detection requirements. However, there are two important differences: first, the algorithm requires only a small number only a small number of frames of the observed sequence to operate. Second, the algorithm uses a max-product algorithm [23] instead of using the common Viterbi algorithm,. The max-product method is a message passing algorithm that makes direct use of the graph structure in constructing and passing messages, and is also very simple to implement.

The proposed algorithm consists of two different blocks: on-line temporal alignment and spatial alignment.

A. On-line temporal alignment

On-line temporal alignment, or synchronization, consists of associating each newly acquired image (in the observed sequence) to one of the frames in the reference sequence. That is, a single frame is only estimated for each newly acquired image instead of a frame correspondence function. This task is formulated as a probabilistic labeling problem. A label $x_t \in \{1, \dots, n_r\}$ refers the frame number in reference sequence associated to the t^{th} newly acquired frame. Hence, the label x_{t-l} is inferred using fixed-lag smoothing [8] on a

hidden Markov model that only considers a small number of frames of the observed sequence as follows:

$$x_{t-l}^* = \operatorname{argmax}_{x_{t-l} \in \Omega_t} p(x_{t-l} | \mathbf{y}_{t-L:t}), \quad (1)$$

where x_{t-l}^* is the inferred label at time t for the frame recorded l time units ago. Ω_t is the set of possible labels, $l \geq 0$ is the lag or delay of the system, $\mathbf{y}_{t-L:t}$ are the observations from the $(t-L)^{th}$ to t^{th} frame in the observed sequence and $L+1 \geq l$ is the total number of observations available used for inferring the current label x_{t-l} .

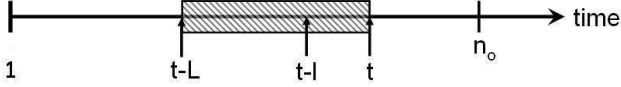


Fig. 3. Representation of fixed-lag smoothing inference. Label x_{t-l} is estimated at time t using only $L+1$ frames in the observed sequence.

The aim of the fixed-lag smoothing is estimating the label x_{t-l} that must show 'similar content' among the $L+1$ frames in the observed sequence (Fig. 3). Hence, the most likely adjacent labels of x_{t-l} , $\mathbf{x}_{t-L:t}$, and the observed frames $\mathbf{y}_{t-L:t}$ must also show 'similar content'. Hence, the $p(x_{t-l} | \mathbf{y}_{t-L:t})$ is formulated to maximize the frame similarity given the most likely temporal mapping as follows:

$$p(x_{t-l} | \mathbf{y}_{t-L:t}) = \max_{\mathbf{x}_{t-L:t} \setminus x_{t-l}} p(\mathbf{x}_{t-L:t} | \mathbf{y}_{t-L:t}), \quad (2)$$

where $\mathbf{x}_{t-L:t} = [x_{t-L}, \dots, x_t]$ is the temporal mapping between the reference sequence and $t-L:t$ frames in the observed sequence, $\mathbf{x}_{t-L:t} \setminus x_{t-l}$ considers all variables except to x_{t-l} , and $p(\mathbf{x}_{t-L:t} | \mathbf{y}_{t-L:t})$ measures the temporal correspondence between the observed frames and a reference sequence. Furthermore, the posterior probability density in Eq. (2) is decoupled as follows:

$$p(\mathbf{x}_{t-L:t} | \mathbf{y}_{t-L:t}) \propto p(\mathbf{y}_{t-L:t} | \mathbf{x}_{t-L:t}) p(\mathbf{x}_{t-L:t}), \quad (3)$$

where $p(\mathbf{y}_{t-L:t} | \mathbf{x}_{t-L:t})$ and $p(\mathbf{x}_{t-L:t})$ are the observation likelihood and the *prior*, respectively. This prior $p(\mathbf{x}_{t-L:t})$ favors only labellings that satisfies some assumptions (e.g., the vehicles can stop independently), whereas the observation likelihood $p(\mathbf{y}_{t-L:t} | \mathbf{x}_{t-L:t})$ measures the similarity between a pair of videos given a temporal mapping $\mathbf{x}_{t-L:t}$. Finally, the max-product algorithm is used to infer the label x_{t-l}^* as follows:

$$x_{t-l}^* = \operatorname{argmax}_{x_{t-l} \in \Omega_t} \max_{\mathbf{x}_{t-L:t} \setminus x_{t-l}} p(\mathbf{y}_{t-L:t} | \mathbf{x}_{t-L:t}) p(\mathbf{x}_{t-L:t}). \quad (4)$$

For simplicity, the prior $p(\mathbf{x}_{t-L:t})$ and the observation likelihood $p(\mathbf{y}_{t-L:t} | \mathbf{x}_{t-L:t})$ are factorized as follows:

$$p(\mathbf{x}_{t-L:t}) = P(x_{t-L}) \prod_{k=t-L}^{t-1} p(x_{k+1} | x_k) \quad (5)$$

$$p(\mathbf{y}_{t-L:t} | \mathbf{x}_{t-L:t}) = \prod_{k=t-L}^t p(y_k | x_k), \quad (6)$$

under the assumption that the transition and the observation probabilities are conditionally independent given the previous and current label values, respectively. $P(x_{t-L})$ gives the same probability to all labels Ω_t to avoid the propagation of possible errors in temporal assignments.

The intended meaning of $p(x_{k+1} | x_k)$ is that vehicles do not go backward, that is, they move always forward or at most stop for some time. Therefore, labels x_t must increase monotonically as follows:

$$p(x_{k+1} | x_k) = \begin{cases} \beta & \text{if } x_{k+1} \geq x_k \\ 0 & \text{otherwise} \end{cases}, \quad (7)$$

where β is a constant that gives the same importance to all label configurations $\mathbf{x}_{t-L:t}$ satisfying the constraint in Eq. (7). That *prior* $p(x_{k+1} | x_k)$ does not restrict the vehicle speed, but they can vary independently.

The intended meaning of $p(y_k | x_k)$ is to measure a frame similarity given a pair of frames and is defined as follows:

$$p(y_k | x_k) = \Phi(f(y_k, y_{x_k}); \mu_y, \sigma_y^2), \quad (8)$$

where y_k and y_{x_k} are the image descriptors of the k^{th} and the x_k^{th} frames in the observed and reference sequence respectively, $f(y_k, y_{x_k}^{i,j})$ is a similarity measure between both descriptors, and $\Phi(v; \mu_y, \sigma_y^2)$ denotes the evaluation of the Gaussian pdf $\mathcal{N}(\mu, \sigma_y^2)$ at v , being μ_y and σ_y the mean and variance of the similarity measure $f(\cdot, \cdot)$.

B. Spatial alignment

Spatial alignment consists of estimating a geometric transformation that relates the image coordinates of a pair of corresponding frames. For any such pair, the cameras are assumed to be at the same position but their 3D orientation (pose) may be different because of trajectory differences, and acceleration, braking and road surface irregularities affecting the yaw and pitch angles, respectively. Hence, the geometric transform existing between two corresponding frames is an special class of homography, the conjugate rotation $H = KRK^{-1}$, being $K = \text{diag}(f, f, 1)$ and f the focal length of the camera in pixels. It is important to bear in mind that, despite this notation does not express it for sake of simplicity, this transformation is not constant along the whole sequence. The transformation changes for every pair of corresponding frames thus making difficult the synchronization task. The rotation matrix R expresses the relative orientation of the cameras for one pair of corresponding frames, and it is parametrized by the Euler angles $\Omega = (\Omega_x, \Omega_y, \Omega_z)$ (pitch, yaw and roll respectively). Furthermore, the transformation modeled by H is approximated using a quadratic motion model as follows [24]:

$$\mathcal{W}(\mathbf{x}; \Omega) = \begin{bmatrix} -\frac{xy}{f} & f + \frac{x^2}{f} & -y \\ -f - \frac{y^2}{f} & \frac{xy}{f} & x \end{bmatrix} \begin{bmatrix} \Omega_x \\ \Omega_y \\ \Omega_z \end{bmatrix}. \quad (9)$$

In this way Ω is estimated minimizing the sum of squared differences by means of the additive forward implementation of the Lukas–Kanade algorithm [25] as follows:

$$\Omega^* = \operatorname{argmin}_{\Omega} \left(\sum_{\mathbf{x}} [S_{x_t}^r(\mathbf{x} + \mathcal{W}(\mathbf{x}; \Omega)) - S_t^o(\mathbf{x})]^2 \right), \quad (10)$$

where $S_{x_t}^r$ is the image warped onto the image coordinates of S_t^o . Ω is iteratively estimated in a coarse-to-fine manner. For a detailed description we refer the reader to [25].

IV. ROAD DETECTION BASED ON VIDEO ALIGNMENT

In this section, a novel road detection algorithm based on on-line video alignment is proposed. The algorithm consists of two stages: on-line video alignment and refinement. Thus, the proposed algorithm combines the robustness of video alignment to provide road segmentations despite lighting conditions and the accuracy of a pixel-level refinement process. The first stage relates frame- and pixel-wise two video sequences and transfers road knowledge from the first sequence to the current ride. Further, the algorithm improves robustness against lighting variations and shadows by using a shadowless feature space. The second stage is a based on dynamic background subtraction to remove objects in the observed sequence. This refinement consists of analyzing road regions based on the image dissimilarities between both rides. Thus, the process assumes the stored sequence is recorded with the absence or low-density traffic.

A. On-line video alignment for road detection

The first stage of the algorithm consists of applying the on-line video alignment method (Sect. III). Moreover, robustness against lighting variations and shadows is improved by using an illuminant-invariant feature space [26] to perform image comparisons (Eq. (8)). This illuminant-invariant space minimizes the influence of lighting variations under the assumption of *Lambertian* surfaces imaged by a three fairly narrow-band sensor under approximately Planckian light sources [26]. As shown in Fig. 4, the characterization process (i.e., converting an RGB image onto the shadow-less feature space) consists of projecting the $\{\log(R/G), \log(B/G)\}$ pixel values of the image onto the direction orthogonal to the lighting change lines, *invariant-direction* θ . This direction is device dependent and can be estimated off-line using the calibration procedure of [26].

In practice, image descriptors y_* are computed as follows. First, the image converted onto the shadow-less feature space shown in Fig. 5 is smoothed using a Gaussian kernel with σ and downsampled along each axis at $1/16^{\text{th}}$ of the original resolution. Then, partial derivatives are computed setting the gradient magnitude at each pixel equal to zero if it is less than 5% of the maximum. The reason to employ a low threshold

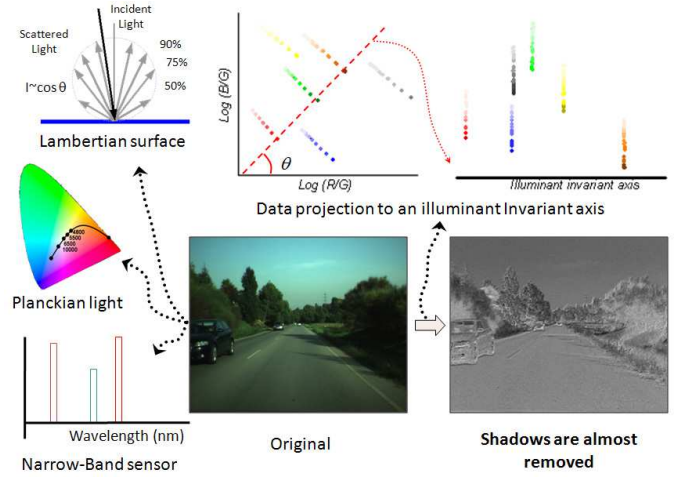


Fig. 4. An illuminant-invariant image is obtained under the assumptions of Planckian light, Lambertian surface and narrow-band sensors. This image is almost shadow free.

of the gradient magnitude instead of the intensity value itself is to reinforce to the lighting invariance conditions. Finally, all the partial derivatives are stacked into a column vector that is normalized to unit norm. Then, the similarity measure $f(\cdot, \cdot)$ is defined as the maximum of the inner product between descriptors among different horizontal and vertical translations of the smoothed downsampled input image, which are set up to 2 pixels. Thus, $\Phi(v; \mu_y, \sigma_y^2)$ is related to the closest coincidence angle between two descriptor vectors, whose μ_y and σ_y are set empirically to 1 and 0.5. That maximum makes the similarity measure in Eq. (8) invariant, to some extent, to slight rotations and translations between the reference and observed frames. These dissimilarities are unavoidable because, of course, the vehicle will not follow exactly the same trajectory in the two rides.

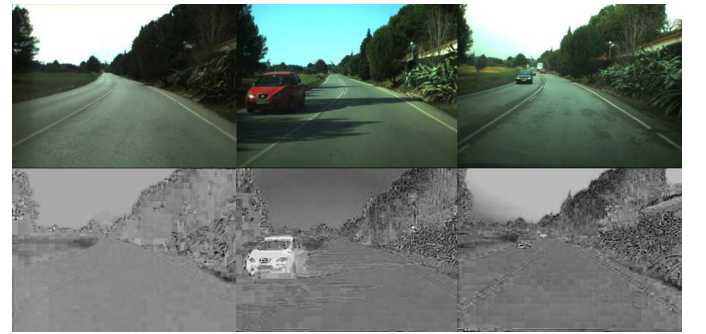


Fig. 5. Illuminant-invariant examples of images acquired approximately at the same position under different lighting conditions.

Figure 6 shows the synchronization benefits of using the illuminant-invariant feature space. Further, quantitative evaluation results in a lower average synchronization error when the illuminant-invariant feature space is used (1.05 ± 0.8791 against 1.75 ± 0.87). From this results we can conclude that using the illuminant-invariant representation improves the accuracy of the algorithm to discriminate corresponding frames.

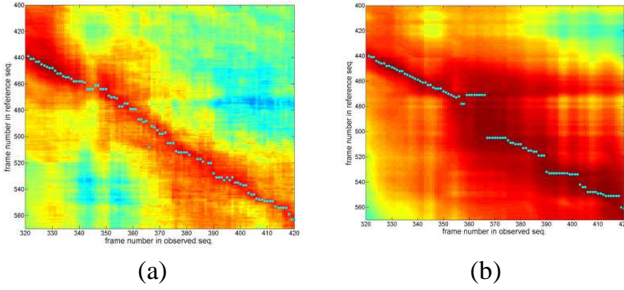


Fig. 6. Using an illuminant-invariant feature space improves the performance of the on-line synchronization algorithm. Images show on-line synchronization (filled circles) on a background inversely proportional to frame similarity in Eq. (8). a) Synchronization results using the illuminant-invariant feature space. b) Synchronization results using gray level images.

B. Road refinement

The result of the video alignment stage is a pair of corresponding frames (t, x_t) and its relative geometric transformation Ω_t that relates them pixel-wise. Hence, road regions \mathcal{M}_t delimiting the road surface in the observed sequence at time t are obtained by warping the road segmentation of the corresponding frame in the reference sequence \mathcal{M}_{x_t} using Ω_t as follows:

$$\mathcal{M}_t = \mathcal{M}_{x_t}(\mathbf{x} + \mathcal{W}(\mathbf{x}; \Omega_t)). \quad (11)$$

However, the transferred road regions \mathcal{M}_t is a rough approximation of the free-road surface due to the observed sequence may show different objects (e.g., vehicles, pedestrian). Therefore, the refinement algorithm removes regions that contain those objects in the observed frames and are within the transferred road regions. This assumption is because the detected objects are claimed to the observed sequence due to the fact that the first ride is recorded with the absence or low-density traffic. Therefore, a dynamic background subtraction is proposed to detect objects spotting differences between a corresponding frame pair.

In particular, the dynamic background subtraction is computed as follows (Fig. 7). First, the corresponding frame in the reference sequence $S_{x_t}^r$ is warped to the image coordinates of the observed frame, $S_{x_t}^{r,w}$. Then, the intensity of a pair of corresponding frames are subtracted pixel-wise as $S_{x_t}^{r,w} - S_t^o$, being S_t^o the observed frame at time t . Hence, the subtraction allows to spot differences of potential interest that are considered as objects present in the observed sequence. Hence, the detected regions are considered forward objects. Then, the absolute value of the pixel subtraction, $|S_{x_t}^{r,w} - S_t^o|$, is binarized using automatic thresholding techniques [27], and the possible holes in the binary regions are filled using mathematical morphology. Fig. 7 illustrates an example of refinement procedure to remove regions that contains vehicles and are within the transferred road.

V. EXPERIMENTS

In this section, two different experiments are conducted to validate the proposed algorithm on different sequences

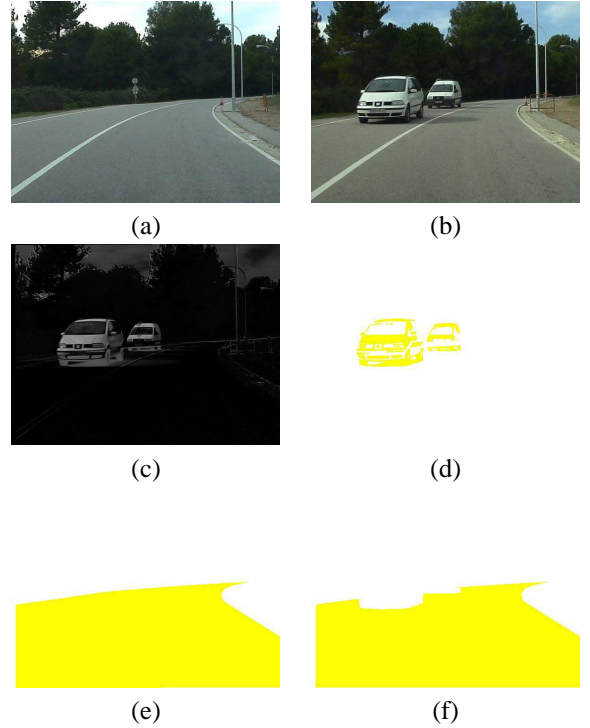


Fig. 7. Refinement shows the corresponding frame from the database (a) is aligned with the input frame (b). The difference between them (c) is used to detect foreground objects (d). The known road surface of the corresponding frame is transferred to the input image (e), and finally, the foreground vehicles are removed (f).

acquired with a forward-facing camera attached at the wind-screen. The goal in the first experiment is detecting free-road areas ahead of a vehicle. The second experiment consist of applying the algorithm to automatically generate ground-truth to evaluate the performance of road detection algorithms.

A. Datasets

Experiments are conducted on three different scenarios: 'street', 'back-road' and 'campus'. The first two scenarios consist of three video sequence pairs; whereas 'campus' scenarios only one video sequence pair. The following pairs, 'Street-1' and 'street-2' provided by [28], and 'back-road-1' and 'back-road-2', are three video sequences recorded following the same route, one reference and two observed sequences, to demonstrate the robustness of inferring free-drivable areas under different lighting conditions. The observed sequence in 'back-road-1', 'back-road-3', 'street-2', 'street-3' and reference in 'campus' are recorded at noon in a sunny day under the presence of shadows on the road surface; whereas the observed sequence in 'street-1' is recorded in a sunny day at morning under the presence of shining road surface. The rest of sequences do not contain shadows because the reference sequences except in 'campus' are recorded in a cloudy day whereas the observed sequence in 'campus' was acquired during the sunset and the observed sequence in 'back-road-3' is acquired in a cloudy day with a wet road surface. Furthermore, both sequences in 'street-1' and 'street-3' scenario are free of vehicles in contrast with 'back-road-

1' and 'back-road-1' that contain vehicles in both sequences. The rest of sequence pairs deal with the presence of vehicles in the observed sequence. The number of frames in observed sequence differs from the reference sequences due to differences in the trajectory and speed of the vehicle. Table I summarizes the main characteristics of each scenario and sequences to demonstrate the variability of the experiments. All sequences are recorded at the same frame rate. The road regions of the reference and observed sequence in all scenarios are manually delineated.

B. Performance Evaluation

Quantitative evaluations are provided using pixel-based measures defined in a contingency table (Table II). The entries of this table are defined as follows: TP is the number of correctly labelled road pixels, TN is the number of non-road pixels detected, FP is the number of non-road pixels classified as road pixels and FN is the number of road pixels erroneously marked as non-road. Further, using the entries in the contingency table, the following error measures are computed: accuracy, sensitivity, specificity and quality, see Table III. Each of these measures provides different insight of the results. Accuracy provides information about the fraction of classifications that are correct. Specificity measures the proportion of true negatives (i.e., background pixels) which are correctly identified. Sensitivity, or recall, is the ratio of detecting true positives (i.e., road pixels). Quality is related to the completeness of the extracted data as well as its correctness. All these measures range from 0 to 1, from worst to perfect.

TABLE II

THE CONTINGENCY TABLE. ALGORITHMS ARE EVALUATED BASED ON THE NUMBER OF PIXELS CORRECTLY AND INCORRECTLY CLASSIFIED. SEE TEXT FOR ENTRIES DEFINITION.

Contingency Table		Ground-truth	
		Non-Road	Road
Result	Non-Road	TN	FN
	Road	FP	TP

TABLE III

PIXEL-WISE MEASURES USED TO EVALUATE THE PERFORMANCE OF THE DIFFERENT DETECTION ALGORITHMS. THESE MEASURES ARE DEFINED USING THE ENTRIES OF THE CONTINGENCY TABLE (TABLE II).

Measure	Definition
Quality	$\hat{g} = \frac{TP}{TP+FP+FN}$
Accuracy	$ACC = \frac{TP+TN}{TP+FP+FN+TN}$
Sensitivity	$TPR = \frac{TP}{TP+FN}$
Specificity	$SPC = \frac{TN}{FP+TN}$

C. On-line Road Detection Results

In the first experiment, the significance of including the refinement stage is evaluated comparing the performance before

and after the refinement stage. A summary of quantitative evaluations is listed Table IV and some example results are shown in Fig. 8 and Fig. 9. As shown, road areas are properly transferred from the reference sequence to observed sequence. Specifically, Fig. 9 shows the robustness of transferring the same *road prior* to different observed sequences under different lighting conditions. That is, the proposed algorithm deals with the presence of shadows and a wet road surface (the first five rows in Fig. 9c-d and Fig. 9e-f, respectively), and shining road surface (the last four rows in Fig. 9c-f). Errors (shown in red and green in Fig. 8) and Fig. 9) are mainly located at the road boundary mainly due to the ambiguity of manually delimiting the road boundaries. Furthermore, the refinement step handles correctly the presence of vehicles cropping properly the transferred road region. That step increases the performance in all four measures as shown in Table IV. In addition, Fig. 9c-f shows the benefits (larger discriminative power) of including the refinement step when other vehicles are present in the scene. This is mainly due to the fact that the refinement stage removes on-coming, incoming or parked vehicles. From these results we can conclude that the proposed algorithm is able to recover road areas despite different lighting conditions, e.g., shadows, wet and shining surface, and the presence of vehicles with different size, colors and shapes in the scene. Finally, as shown in the second row of Fig. 9a-b, the proposed algorithm also handles the presence of vehicles (low-dense traffic) in the reference sequence reducing the accuracy since the with vehicles are not transferred. This is also reinforced quantitatively in Table IV where the performance in 'back-road-1' and 'back-road-2' is comparable to the performance of other pairs of sequences.

An inherent limitation of the method is the delay before obtaining the results. This delay is set exactly to 200ms (i.e., 5 frames at 25fps). However, this is a minor limitation if a high frame-rate camera is provided.

D. Off-line road detection: Automatic Ground-truthing

Ground-truth data is a must for the quantitative assessment and comparison of detection/segmentation algorithms. In the context of road detection, the manual annotation of the road regions on sequences hundreds or thousands frames long is very time consuming and prone to error because of the human operator's attention drop off. The required effort is even higher for works that claim to be robust to different lighting conditions like [29], since for one same track there are several sequences that must be manually annotated. Thus, automatic generation of ground-truth for evaluating road detection algorithms is a problem of interest in itself.

The proposed algorithm for automatic ground-truthing transfers the manual annotation in one sequence to another when they are completely recorded. Example results are shown in Fig. 10. More results in video format can be viewed at <http://www.cvc.uab.es/~fdiego/RoadSegmentation/>. These results suggest that the learned ground-truth in the reference sequence is correctly transferred to the observed one. As shown in Fig. 10d, errors are mainly located at the road boundaries. However, these errors are due to the inherent

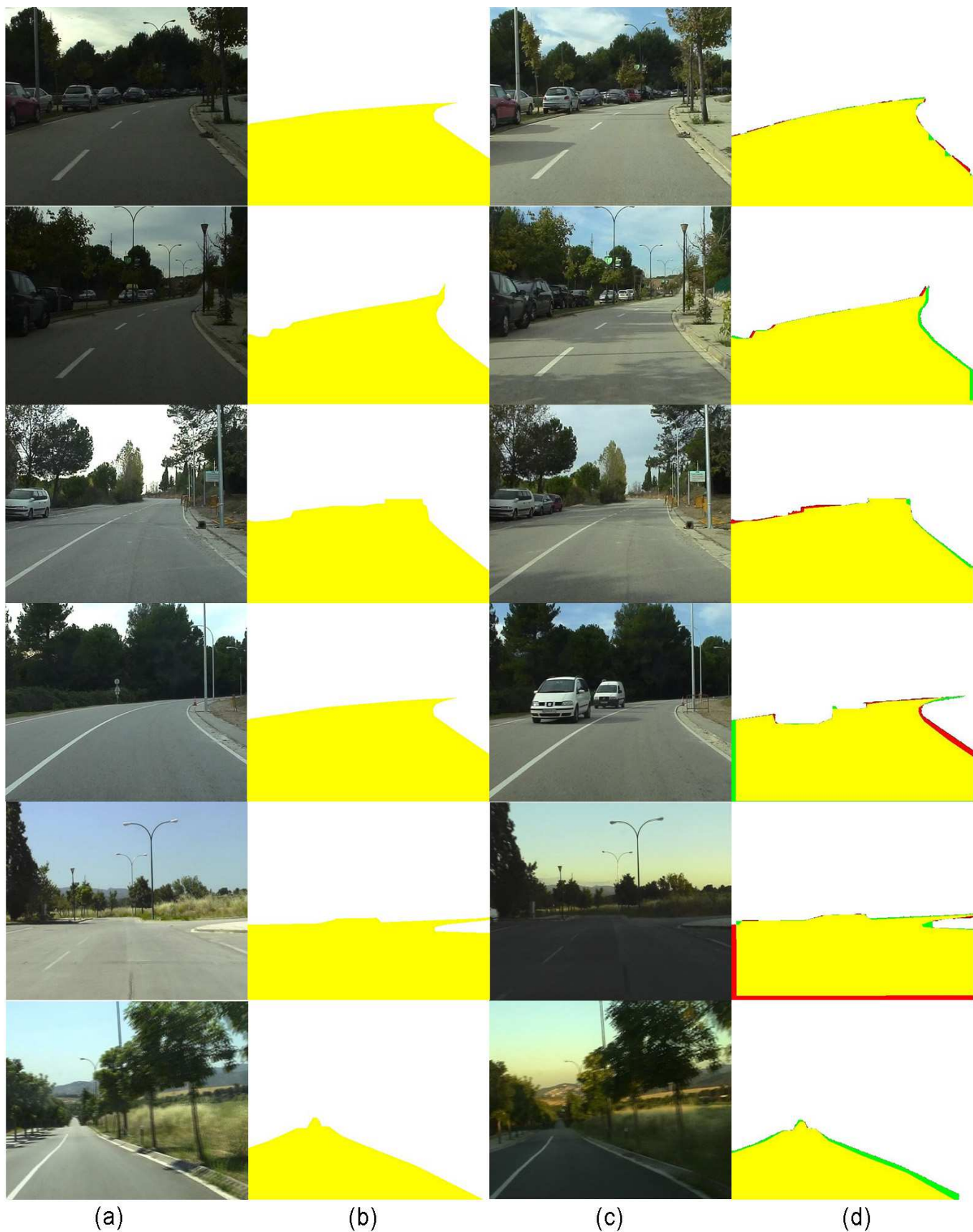


Fig. 8. Example results of the proposed road detection algorithm for different scenarios. The frame from the reference sequence (a) is aligned with the input frame (c). Learned road regions (b) combined with the refinement stage are used to generate the final result (d). The color code in the image is as follow: true positives are in yellow; true negatives are in white; false positives are in read and, false negatives are in green, with respect to a road/non-road classification. More results, in video format, can be viewed at <http://www.cvc.uab.es/~fdiego/RoadSegmentation/>.

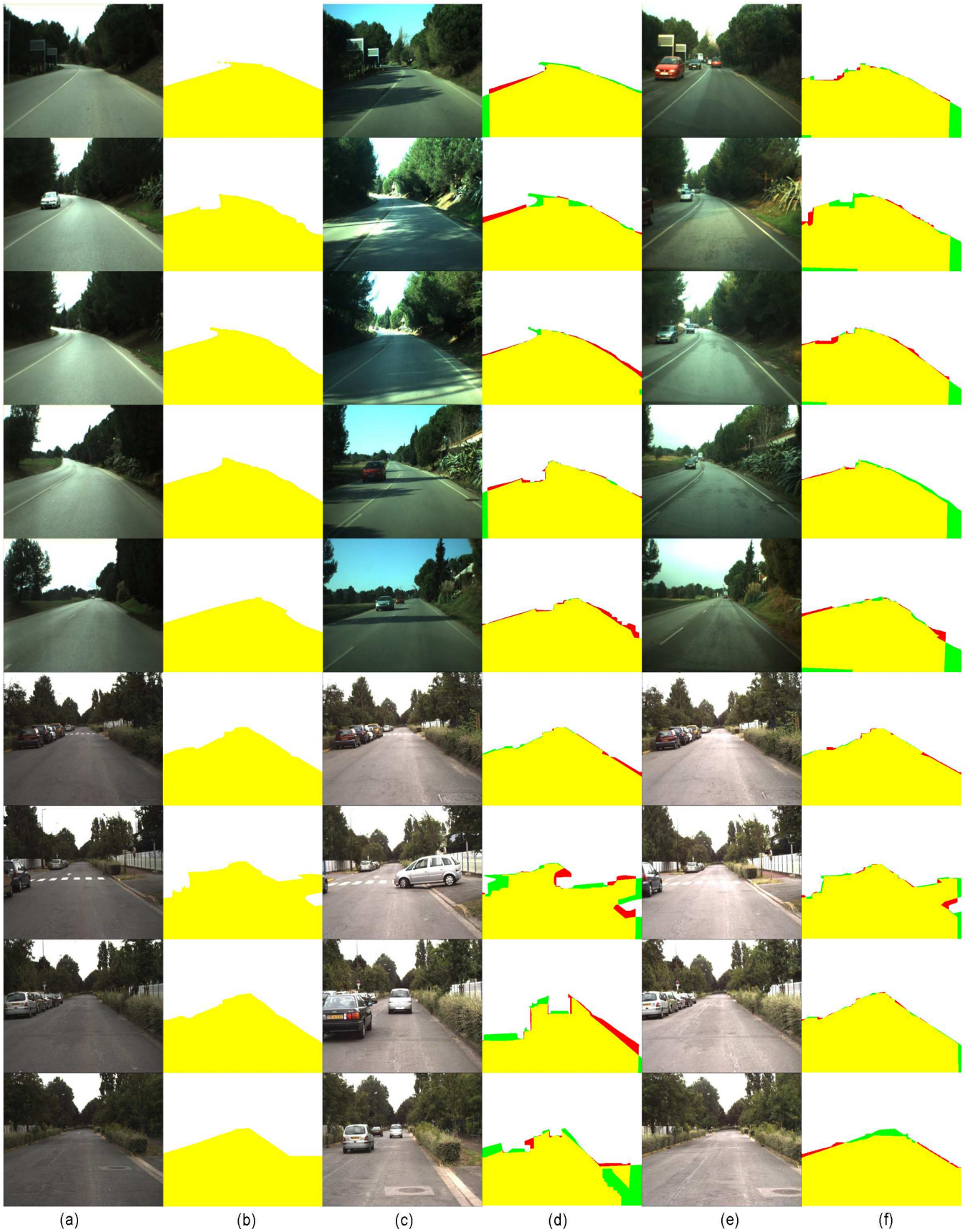


Fig. 9. Example results of the proposed road detection algorithm for two different scenarios driven, at least, 3 times. The same frame from the reference sequence (a) is aligned with the input frames (c) and (e) under different lighting conditions. Learned road regions (b) combined with the refinement stage are used to generate the final results (d) and (f). The color code in the image is as follow: true positives are in yellow; true negatives are in white; false positives are in read and, false negatives are in green, with respect to a road/non-road classification.

TABLE I
DESCRIPTION OF THE MAIN CHARACTERISTICS OF THE SCENARIOS AND SEQUENCES.

Scenario	Sequence	Recording time	weather	Shadows	Vehicles	length
Back-Road-1	Observed	noon	sunny	yes	yes	714
	Reference	morning	cloudy	no	yes	948
Back-Road-2	Observed	noon	cloudy	wet	yes	402
	Reference	morning	cloudy	no	yes	948
Street-1 [11]	Observed	noon	sunny	yes	yes	210
	Reference	noon	cloudy	no	no	239
Street-2 [11]	Observed	morning	sunny	shining	no	260
	Reference	noon	cloudy	no	no	239
Street-3	Observed	noon	sunny	yes	no	520
	Reference	afternoon	cloudy	no	no	627
Back-Road-3	Observed	noon	sunny	yes	yes	1318
	Reference	afternoon	cloudy	no	no	1459
Campus	Observed	sunset	sunny	no	yes	600
	Reference	noon	sunny	yes	no	816

TABLE IV
AVERAGE PERFORMANCE OF THE PROPOSED ROAD DETECTION ALGORITHM OVER ALL THE CORRESPONDING FRAMES.

Scenario	Refinement	\hat{g}	SPC	TPR	ACC
Back-Road-1	without	0.9637 ± 0.03792	0.9724 ± 0.0285	0.9528 ± 0.05489	0.9640 ± 0.0400
	within	0.9680 ± 0.0304	0.9819 ± 0.0111	0.9504 ± 0.0624	0.9760 ± 0.0163
Back-Road-2	without	0.9425 ± 0.02364	0.9688 ± 0.0238	0.9105 ± 0.0312	0.9597 ± 0.0308
	within	0.9467 ± 0.0193	0.9834 ± 0.0138	0.9018 ± 0.0374	0.97793 ± 0.0187
Street-1 [11]	without	0.9415 ± 0.0380	0.9348 ± 0.0507	0.9556 ± 0.0438	0.9057 ± 0.0795
	within	0.9495 ± 0.0351	0.96001 ± 0.0343	0.9356 ± 0.0617	0.9367 ± 0.0601
Street-2 [11]	without	0.9828 ± 0.0194	0.9897 ± 0.0125	0.9744 ± 0.0386	0.9878 ± 0.0132
	within	0.9846 ± 0.0190	0.9903 ± 0.0094	0.9778 ± 0.0388	0.9885 ± 0.0102
Street-3	without	0.9691 ± 0.0117	0.9914 ± 0.0044	0.9807 ± 0.0099	0.9869 ± 0.0051
	within	0.9817 ± 0.0092	0.9936 ± 0.0039	0.9904 ± 0.0092	0.9923 ± 0.0039
Back-Road-3	without	0.9559 ± 0.0546	0.9869 ± 0.0258	0.9728 ± 0.0236	0.9802 ± 0.0262
	within	0.9626 ± 0.0543	0.9909 ± 0.0244	0.9746 ± 0.0260	0.9832 ± 0.0262
Campus	without	0.9127 ± 0.0890	0.9897 ± 0.0075	0.9403 ± 0.0842	0.9766 ± 0.0219
	within	0.9262 ± 0.0794	0.9909 ± 0.0078	0.9501 ± 0.0713	0.9794 ± 0.0236

boundary ambiguity when the images are manually segmented by an human operator.

Quantitative evaluations are summarized in Table V. Two different evaluations are conducted on 'parking' scenario to demonstrate the capacity of generating accurate ground-truth using any traffic-free sequence as a reference sequence. The former keeps the same nomenclature of sequences whereas the latter interchanges the reference sequence as an observed sequence and vice versa. The averaged performance over all the corresponding frames is shown in Table V. Labelling an image takes 30 seconds in average time so using the algorithm on 'parking' sequences saves 4.3 and 3.7 hours, respectively. Small differences are due to the different number of frames in each video sequence. The highest performance is achieved when the largest video sequence is used as reference. The main reason is that the algorithm does not interpolate the information between frames. Thus, the large amount information available as reference, the highest accuracy in the registration process. However, this is a minor drawback since the reference sequence could be recorded driving at a lower speed or at a higher frame-rate.

An inherent limitation of the method is the presence of

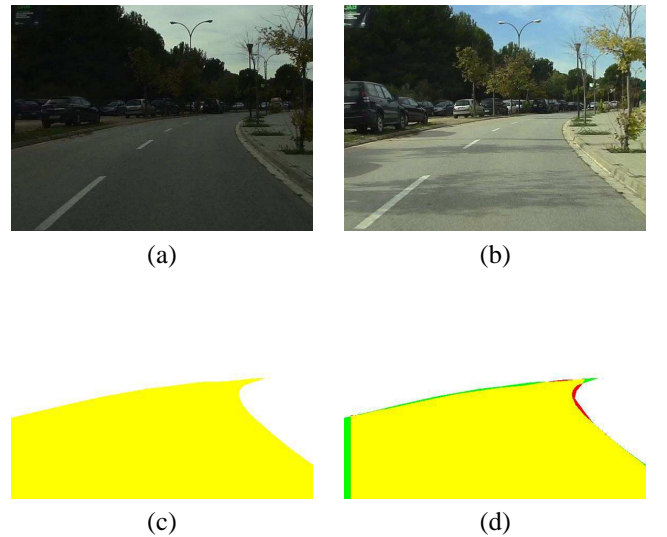


Fig. 10. Example results of the proposed automatic ground-truthing algorithm. The frame from the learned sequence (a) is aligned with the new acquired frame (b). The reference ground-truth (c) is used to generate the output ground-truth (d). Yellow color refers to true positive pixels. White means true negative pixels, red false positives and green false negatives.

TABLE V
PERFORMANCE OF THE GROUND-TRUTHING ALGORITHM CONDUCTED ON 'PARKING' SCENARIO.

	\bar{g}	SPC	TPR	ACC
Reference Seq. as reference	0.9784 ± 0.0103	0.9928 ± 0.0057	0.9882 ± 0.0095	0.9909 ± 0.0044
Observed Seq. as reference	0.9748 ± 0.0535	0.9914 ± 0.0068	0.9867 ± 0.0538	0.9894 ± 0.0218

moving vehicles in the reference sequence. However, it is a minor drawback because the road regions occluded by vehicles can be interpolated according to the available road boundaries. Further, the algorithm can be used in semi-supervised mode. That is, the ground-truth is automatically generated and shown to the operator for validation.

VI. CONCLUSION

In this paper on-line video alignment has been introduced for road detection. The key idea of the algorithm is to exploit similarities occurred when a vehicle follows the same route more than once. Hence, road knowledge is learnt in a first ride and then, this knowledge is used to infer road areas in subsequent rides. Furthermore, a dynamic background subtraction is proposed to handle correctly the presence of vehicles cropping properly the inferred road region. Thus, the algorithm combines the robustness against local lighting variations of video alignment with the accuracy at pixel-level provided by the refinement step. Experiments are conducted on different image sequences taken at different day time on real-world driving scenarios. From qualitative and quantitative results, we can conclude that the proposed algorithm is suitable for detecting the road despite varying lighting conditions (i.e., shadows and different daytime) and the presence of other vehicles in the scene.

REFERENCES

- [1] C. Thorpe, M. Hebert, T. Kanade, and S. Shafer, "Vision and navigation for the carnegie-mellon navlab," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 10, no. 3, pp. 362 – 373, May 1988.
- [2] A. Lookingbill, J. Rogers, D. Lieb, J. Curry, and S. Thrun, "Reverse optical flow for self-supervised adaptive autonomous robot navigation," *International Journal of Computer Vision (IJCV)*, vol. 74, no. 3, pp. 287–302, 2007.
- [3] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 32, pp. 1239–1258, 2010.
- [4] C. Rasmussen, "Grouping dominant orientations for ill-structured road following," in *CVPR'04: Procs. of the IEEE on Computer Vision and Pattern Recognition*, Washington, DC, 2004, pp. 470–477.
- [5] M.A. Sotelo, F.J. Rodriguez, and L. Magdalena, "Virtuous: vision-based road transportation for unmanned operation on urban-like scenarios," *IEEE Trans. Intelligent Transportation Systems (ITS)*, vol. 5, no. 2, pp. 69 – 83, June 2004.
- [6] C. Tan, T. Hong, T. Chang, and M. Shneier, "Color model-based real-time learning for road following," in *ITSC'06: Procs. IEEE Intl. Conf. on Intel. Transp. Systems*, 2006, pp. 939–944.
- [7] Y. Caspi and M. Irani, "Spatio-temporal alignment of sequences," *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 24, no. 11, pp. 1409–1424, 2002.
- [8] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson Education, 2003.
- [9] P. Lombardi, M. Zanin, and S. Messelodi, "Switching models for vision-based on-board road detection," in *ITSC'05: Procs. IEEE Intl. Conf. on Intel. Transp. Systems*, Vienna, Austria, 2005, pp. 67 – 72.
- [10] Y. He, H. Wang, and B. Zhang, "Color-based road detection in urban traffic scenes," *IEEE Trans. Intelligent Transportation Systems (ITS)*, vol. 5, no. 24, pp. 309 – 318, 2004.
- [11] Hui Kong, Jean Yves Audibert, and Jean Ponce, "General road detection from a single image," *IEEE Transactions on Image Processing (TIP)*, vol. 19, no. 8, pp. 2211 – 2220, 2010.
- [12] C. Rotaru, T. Graf, and J. Zhang, "Color image segmentation in HSI space for automotive applications," *Journal of Real-Time Image Processing*, pp. 1164–1173, 2008.
- [13] T. Michalke, R. Kastner, M. Herbert, J. Fritsch, and C. Goerick, "Adaptive multi-cue fusion for robust detection of unmarked inner-city streets," in *IV'09: Procs. of the IEEE Intel. Vehicles Symposium*, June 2009, pp. 1 – 8.
- [14] M.A. Sotelo, F. Rodriguez, L.M. Magdalena, L. Bergasa, and L. Boquete, "A color vision-based lane tracking system for autonomous driving in unmarked roads," *Autonomous Robots*, vol. 16, no. 1, pp. 95–116, 2004.
- [15] Flavio L.C. Padua, Rodrigo L. Carceroni, Geraldo A.M.R. Santos, and Kiriakos N. Kutulakos, "Linear sequence-to-sequence alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 32, pp. 304–320, 2010.
- [16] M. Singh, I. Cheng, M. Mandal, and A. Basu, "Optimization of symmetric transfer error for sub-frame video synchronization," in *ECCV'08: Proc. European Conference on Computer Vision*, 2008, pp. 554–567.
- [17] Philip A. Tresadern and Ian D. Reid, "Video synchronization from human motion using rank constraints," *Computer Vision and Image Understanding (CVIU)*, vol. 113, no. 8, pp. 891 – 906, 2009.
- [18] Avinash Ravichandran and René Vidal, "Video registration using dynamic textures," *IEEE Trans on Pattern Analysis and Machine Intelligence (PAMI)*, 2011.
- [19] L. Wolf and A. Zomet, "Wide baseline matching between unsynchronized video sequences," *International Journal of Computer Vision (IJCV)*, vol. 68, no. 1, pp. 43–52, 2006.
- [20] Yaron Ukrainitz and Michal Irani, "Aligning sequences and actions by maximizing space-time correlations," in *European Conf. on Computer Vision*, 2006, vol. 3953 of *Lecture Notes in Computer Science*, pp. 538–550, Springer.
- [21] P. Sand and S. Teller, "Video matching," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, vol. 22, no. 3, pp. 592–599, 2004.
- [22] F. Diego, D. Ponsa, J. Serrat, and A. M. Lopez, "Video alignment for change detection," *IEEE Transactions on Image Processing (TIP)*, vol. Preprint, no. 99, 2010.
- [23] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.
- [24] M. Irani, "Multi-frame correspondence estimation using subspace constraints," *International Journal of Computer Vision (IJCV)*, vol. 48, no. 3, pp. 173–194, 2002.
- [25] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International Journal of Computer Vision (IJCV)*, vol. 56, no. 3, pp. 221–255, 2004.
- [26] G.D. Finlayson, S.D. Hordley, C. Lu, and M.S. Drew, "On the removal of shadows from images," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 28, no. 1, pp. 59–68, 2006.
- [27] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, January 1979.
- [28] Hui Kong, J.-Y. Audibert, and J. Ponce, "Detecting abandoned objects with a moving camera," *IEEE Transactions on Image Processing (TIP)*, vol. 19, no. 8, pp. 2201 – 2210, 2010.
- [29] J. M. Alvarez, A. M. Lopez, and R. Baldrich, "Illuminant-invariant model-based road segmentation," in *IV'08: Procs. of the IEEE Intel. Vehicles Symposium*, Eindhoven, The Netherlands, June 2008, pp. 1175–1180.